INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

PROYECTO IA

ALVARO JAVIER TALAIGUA ROSARIO

1196969101

INGENIERÍA MECÁNICA

WILSON ARBEY CHAVERRA MEJÍA

8064138

INGENIERÍA MECÁNICA

UNIVERSIDAD DE ANTIOQUIA

MEDELLIN

2023

Home Credit Default Risk

Enlace de la competición de Kaggle:

https://www.kaggle.com/c/home-credit-default-risk/overview

Enlace del video:

https://drive.google.com/file/d/1Wo34lVAX78BWtv9CuZ_20krBK_jEtjGH/view?usp=drive_link

Enlace de librería en Drive:

https://drive.google.com/drive/folders/1HD7O5R3kE7uGDCYop2Vd9Y8JFbbZgzzn?usp=drive_link

El propósito de este informe es ofrecer una panorámica general y un análisis detallado de la competición "Home Credit Default Risk" en Kaggle. En este desafío, los participantes se encuentran ante la tarea de crear modelos de aprendizaje automático que puedan prever la probabilidad de que los clientes de Home Credit no cumplan con sus pagos.

La competición se enmarca en Home Credit, una entidad financiera no bancaria que proporciona servicios de crédito a individuos con historial crediticio limitado o inexistente. La meta de Home Credit es facilitar el acceso a préstamos y servicios financieros a una parte de la población que usualmente queda fuera de los servicios bancarios convencionales. No obstante, la evaluación del riesgo crediticio en este grupo de clientes puede presentar desafíos debido a la escasez de datos crediticios convencionales.

En este informe, examinaremos exhaustivamente el conjunto de datos suministrado para la competición, que engloba una variedad extensa de características e información acerca de los solicitantes de crédito. Estos datos comprenden aspectos demográficos, historiales laborales, registros de pagos anteriores, transacciones financieras previas, entre otros. A partir de esta información, se espera que los participantes desarrollen modelos predictivos capaces de estimar con precisión la probabilidad de que un cliente no cumpla con sus pagos de crédito.

En el análisis exploratorio de datos, vamos a estudiar la distribución de la variable objetivo, identificar posibles desequilibrios de clases y evaluar la calidad e integridad del conjunto de datos. Además, exploraremos las relaciones y patrones dentro de los datos, tanto entre las variables como en relación con la variable objetivo, con el fin de extraer información significativa.

La competición "Home Credit Default Risk" de Kaggle plantea el desafío de crear modelos de aprendizaje automático para anticipar la probabilidad de que un cliente no cumpla con sus pagos de crédito. A continuación, se presenta una exploración descriptiva elemental del conjunto de datos:

Importancia de los archivos:

● **application_train.csv:** Este archivo contiene los datos de entrenamiento principales, que incluyen información sobre los solicitantes de crédito.

● **application_test.csv**: Este archivo contiene los datos de prueba, y se utiliza para evaluar el rendimiento de los modelos predictivos.

● **bureau.csv:** Proporciona datos adicionales sobre los préstamos anteriores de los solicitantes de crédito, si están disponibles.

● **bureau_balance.csv:** Este archivo contiene información mensual sobre el saldo de deuda de los préstamos anteriores del archivo "bureau.csv".

● **previous_application.csv:** Proporciona datos sobre las aplicaciones previas de los solicitantes de crédito en Home Credit.

● **POS_CASH_balance.csv:** Contiene información mensual sobre los saldos de los préstamos anteriores relacionados con créditos en tiendas.

● **credit_card_balance.csv:** Proporciona datos mensuales sobre los saldos de las tarjetas de crédito de los solicitantes de crédito.

● **installments_payments.csv:** Contiene información sobre los pagos mensuales de los préstamos anteriores.

● **HomeCredit_columns_description.csv:** Un archivo que describe el significado de las columnas en los archivos de datos principales.

Carga de datos: Debes cargar los archivos CSV en tu entorno de programación y familiarizarte con los datos que contienen.

Exploración básica de datos:

● Comienza por visualizar algunas filas de los datos para entender la estructura y los tipos de variables.

● Examina la distribución de la variable objetivo (TARGET) para ver si hay un desequilibrio significativo.

● Calcula y analiza las estadísticas descriptivas de las variables numéricas (media, mediana, desviación estándar, etc.).

● Explora las variables categóricas y observa los diferentes valores únicos y su frecuencia.

● Verifica si hay valores faltantes en el dataset y decide cómo manejarlos.

Análisis más detallado:

● Examina las relaciones entre las variables y la variable objetivo mediante gráficos y medidas de correlación.

● Realiza análisis comparativos entre las características de los clientes que incumplen y los que no.

● Investiga si hay variables altamente correlacionadas y considera la posibilidad de eliminar algunas para evitar la multicolinealidad.

La implementación de un modelo de aprendizaje automático conlleva diversos desafíos y consideraciones esenciales. Aquí se presentan algunos de los retos clave y aspectos a tener en cuenta:

Infraestructura y entorno de implementación: Se requiere asegurar una infraestructura adecuada para alojar y ejecutar el modelo, lo que implica considerar servidores, recursos de almacenamiento, capacidad de procesamiento y configuraciones de red. También es crucial tener en cuenta los requisitos de software y las dependencias del modelo.

Escalabilidad: En caso de anticipar un alto volumen de solicitudes y tráfico, es esencial diseñar el sistema de despliegue de manera que sea escalable. Esto implica evaluar la capacidad de respuesta del modelo ante múltiples solicitudes simultáneas y planificar cómo escalar los recursos según sea necesario.

Latencia y tiempo de respuesta: Dependiendo del caso de uso, la latencia y el tiempo de respuesta pueden ser factores críticos. Garantizar que el modelo se ejecute de manera eficiente y rápida es fundamental para asegurar una experiencia de usuario fluida.

Seguridad: La seguridad de los datos y el modelo es de vital importancia. Se deben implementar medidas para proteger los datos sensibles y asegurar que el acceso al

modelo esté restringido y controlado. Además, es crucial considerar la seguridad de las comunicaciones y abordar posibles vulnerabilidades del sistema.

Monitoreo y mantenimiento: Una vez que el modelo esté en producción, es crucial establecer un sistema de monitoreo para supervisar su rendimiento y detectar posibles problemas. Además, se deben tener procedimientos establecidos para realizar actualizaciones, mejoras y mantenimiento periódico del modelo.

Versionado y control de cambios: Es crucial mantener un registro claro de las versiones del modelo y los cambios realizados. Esto facilitará la reproducción de resultados, fomentará la colaboración y permitirá identificar problemas en caso de necesitar revertir a versiones anteriores.

Cumplimiento normativo y ético: Asegurarse de que el despliegue del modelo cumpla con las regulaciones y políticas aplicables es esencial. También se debe tener en cuenta consideraciones éticas, como la equidad y la transparencia en el uso del modelo. Garantizar que el modelo se utilice de manera ética y esté alineado con los estándares normativos contribuye a una implementación responsable y sostenible.

Estos son solo algunos de los retos y consideraciones clave a tener en cuenta al desplegar un modelo de aprendizaje automático. Cada caso puede tener sus propias particularidades y requerimientos adicionales. Se recomienda realizar pruebas exhaustivas y trabajar en colaboración con expertos en el dominio y en infraestructura para garantizar un despliegue exitoso y seguro del modelo.

Para la competición seleccionada de Kaggle "Home Credit Default Risk", hay varios retos y consideraciones específicas que debes tener en cuenta durante el despliegue:

Manejo del desequilibrio de clases: Es probable que el conjunto de datos presente un desequilibrio entre la clase de clientes que incumplen con los pagos y los que no. Esto puede afectar el rendimiento del modelo y la interpretación de las métricas de evaluación. Debes considerar técnicas como el muestreo estratificado, la ponderación de clases o el uso de algoritmos específicos para datos desbalanceados.

Adaptabilidad del modelo: Una vez que hayas desarrollado tu modelo, es importante asegurarte de que pueda adaptarse fácilmente a nuevos datos o actualizaciones. Esto implica diseñar una estructura modular y flexible que permita la incorporación de nuevos datos y el reentrenamiento del modelo de manera eficiente.

Eficiencia computacional: Los modelos de aprendizaje automático pueden ser computacionalmente intensivos, especialmente si se utilizan algoritmos complejos o si el conjunto de datos es grande. Asegúrate de que tu modelo sea eficiente en términos de tiempo de entrenamiento, carga en memoria y predicciones en tiempo real. Esto es esencial para garantizar un despliegue eficaz y rentable del modelo.

Interpretabilidad: La explicabilidad del modelo es crucial, especialmente en un contexto de crédito y riesgo. Debes asegurarte de poder proporcionar explicaciones

claras y comprensibles sobre cómo se toman las decisiones de predicción. Considera el uso de modelos interpretables o técnicas de explicabilidad, como la importancia de características o el análisis de saliencia.

Actualización y mantenimiento: El mundo del crédito y el riesgo está en constante cambio, por lo que es importante mantener tu modelo actualizado. Considera cómo podrías automatizar la actualización del modelo en función de nuevos datos y cómo gestionar los cambios regulatorios o en las políticas de crédito. La capacidad de adaptarse a cambios en el entorno financiero es esencial para garantizar que tu modelo siga siendo preciso y relevante con el tiempo.

Cumplimiento normativo y ético: En el contexto de la competición, es crucial asegurarse de que tu modelo cumpla con todas las regulaciones y políticas aplicables, como las leyes de protección de datos. Además, debes considerar las implicaciones éticas y la equidad en la toma de decisiones crediticias. Garantizar la conformidad con normativas y principios éticos es fundamental para asegurar que el despliegue del modelo sea ético y legalmente sólido.

El preprocesamiento de datos es una etapa crucial para preparar los datos antes de aplicar modelos de aprendizaje automático. Aquí están algunas de las tareas típicas realizadas durante este proceso:

Carga de datos: Utilizamos una biblioteca como Pandas para cargar los datos en un entorno de programación.

Exploración de datos: Analizamos los datos cargados para comprender las columnas y tipos de datos presentes.

Limpieza de datos: Se lleva a cabo la limpieza de datos, que puede incluir la eliminación de duplicados, el manejo de valores faltantes y la corrección de errores.

Codificación de variables categóricas: Convertimos variables categóricas en representaciones numéricas adecuadas para el modelado.

Normalización o escalado: Aseguramos que las variables tengan una escala comparable mediante la normalización o el escalado.

Selección o ingeniería de características: Identificamos las características más relevantes o creamos nuevas características que puedan mejorar el rendimiento del modelo.

En el contexto de modelos supervisados, el proceso de preparación de datos y entrenamiento típicamente implica los siguientes pasos:

Separación de características y variable objetivo: Separamos las características (variables independientes) de la variable objetivo (variable dependiente). Las características son los atributos que se utilizarán para hacer predicciones, y la variable objetivo es la que se está tratando de predecir.

División de datos: Dividimos el conjunto de datos en conjuntos de entrenamiento y prueba. El conjunto de entrenamiento se utiliza para entrenar el modelo, mientras que el conjunto de prueba se reserva para evaluar el rendimiento del modelo en datos no vistos. Esto es crucial para determinar si el modelo generaliza bien a nuevos datos.

Selección de modelo: Escogemos un modelo supervisado adecuado para abordar el problema en cuestión. Ejemplos comunes incluyen regresión logística, árboles de decisión, o redes neuronales, entre otros. La elección del modelo depende de la naturaleza del problema y las características de los datos.

Entrenamiento del modelo: Utilizamos el conjunto de entrenamiento para entrenar el modelo seleccionado. Durante este proceso, el modelo ajusta sus parámetros para aprender la relación entre las características y la variable objetivo.

Predicciones: Una vez entrenado, utilizamos el modelo para realizar predicciones sobre el conjunto de prueba. Estas predicciones nos permiten evaluar cómo se comporta el modelo en datos no vistos.

Evaluación del rendimiento: Evaluamos el rendimiento del modelo utilizando métricas apropiadas. Dependiendo de la naturaleza del problema, podríamos utilizar métricas como precisión, recall o error cuadrático medio. Estas métricas proporcionan información sobre la calidad y la eficacia del modelo en hacer predicciones precisas.

Para la aplicación de modelos no supervisados, como K-means para el agrupamiento, los pasos pueden ser los siguientes:

Selección de características relevantes: Seleccionamos las características relevantes que queremos utilizar en nuestro modelo no supervisado. Estas características son fundamentales para identificar patrones o estructuras en los datos sin tener etiquetas previas.

Filtrado de datos: Filtramos los datos según las características seleccionadas, centrándonos en la información que consideramos relevante para el modelo no supervisado.

Elección de modelo no supervisado: Elegimos un modelo no supervisado apropiado para el problema, como el algoritmo de K-means para realizar el agrupamiento de los datos.

Ajuste del modelo: Ajustamos el modelo utilizando los datos filtrados. Durante este proceso, el algoritmo busca patrones y agrupa los datos en clusters o grupos.

Obtención de etiquetas o grupos: Una vez ajustado, obtenemos las etiquetas o grupos resultantes del modelo. Cada grupo representa una colección de datos que comparten características similares según las dimensiones seleccionadas.

Análisis y visualización: Analizamos y visualizamos los grupos obtenidos para obtener información sobre las estructuras o patrones ocultos en los datos. La visualización de los clusters puede revelar relaciones interesantes entre los datos que pueden no ser evidentes de otra manera.

Después de haber entrenado tanto modelos supervisados como no supervisados, la evaluación y comparación de los resultados son etapas cruciales:

Evaluación de resultados con métricas: Utilizamos métricas apropiadas para evaluar el rendimiento de los modelos. Dependiendo del tipo de problema y del modelo utilizado, las métricas pueden incluir precisión, recall, error cuadrático medio o coeficiente de determinación. Estas métricas proporcionan una medida cuantitativa del desempeño del modelo.

Comparación de modelos: Comparamos los resultados de diferentes modelos o configuraciones para determinar cuál presenta el mejor rendimiento. Esto puede implicar la comparación de métricas clave entre modelos para seleccionar el más adecuado para el problema específico.

Curvas de aprendizaje: Utilizamos curvas de aprendizaje para visualizar el rendimiento del modelo en función del tamaño del conjunto de datos de entrenamiento y la complejidad del modelo. Estas curvas pueden proporcionar información sobre la capacidad del modelo para generalizar a medida que se incrementa la cantidad de datos de entrenamiento o la complejidad del modelo.

Iteración y ajuste: Con base en la evaluación de resultados, iteramos y ajustamos los modelos según sea necesario para mejorar la eficacia y el rendimiento. Esto puede incluir la modificación de hiperparámetros, la selección de características adicionales o la consideración de modelos alternativos.

Los desafíos y consideraciones mencionados son fundamentales para abordar la complejidad de la competición "Home Credit Default Risk" en Kaggle. Aquí se destaca cómo se enfrentaron algunos de estos desafíos:

Datos desbalanceados: La presencia de clases desbalanceadas se abordó mediante técnicas específicas, como el uso de muestreo estratificado, la ponderación de clases o la exploración de algoritmos diseñados para manejar desequilibrios. Estas estrategias ayudaron a mejorar la capacidad del modelo para identificar patrones en ambas clases, incluso cuando la proporción entre ellas era desigual.

Preprocesamiento complejo: Dada la complejidad en la estructura y calidad de los datos, se implementaron técnicas avanzadas de preprocesamiento. Esto incluyó la gestión de valores faltantes, el manejo adecuado de características categóricas y la normalización de variables numéricas para asegurar que los datos fueran aptos para el modelado.

Selección de características relevantes: La gran cantidad de características requirió un análisis detallado para identificar las más relevantes. Se llevaron a cabo técnicas de selección de características para mejorar la precisión del modelo y reducir la complejidad. Esto garantizó que el modelo se centrara en las características más informativas para predecir el incumplimiento crediticio.

Elección del modelo adecuado: Dada la diversidad de modelos supervisados disponibles, se realizó una evaluación exhaustiva para seleccionar el modelo más apropiado. La consideración de modelos como regresión logística, árboles de decisión y redes neuronales permitió identificar aquellos que mejor manejaban el desequilibrio de clases y proporcionaban una precisión sólida en la predicción del incumplimiento crediticio.

Validación cruzada: Se utilizó la validación cruzada para evitar el sobreajuste y evaluar el rendimiento de manera más fiable. Este enfoque implica dividir el conjunto de datos en múltiples subconjuntos, entrenar y evaluar el modelo en diferentes combinaciones de conjuntos de entrenamiento y prueba. Esto proporciona una evaluación más robusta del rendimiento del modelo en datos no vistos.

Ajuste de hiperparámetros: Se llevaron a cabo procesos de ajuste de hiperparámetros para encontrar la combinación óptima que equilibra la complejidad del modelo y su capacidad para generalizar correctamente. Esto se hizo mediante la búsqueda sistemática de valores de hiperparámetros que maximizaran el rendimiento del modelo en los conjuntos de validación.

Interpretación de los resultados: Después de obtener los resultados del modelo, se llevó a cabo una interpretación detallada de las predicciones. Esto implicó entender los factores que influyen en el incumplimiento crediticio y extraer información valiosa para tomar decisiones informadas. La capacidad de interpretar los resultados es crucial para explicar de manera clara y coherente los determinantes del riesgo crediticio a los interesados y partes involucradas

Conclusiones:

El manejo de datos desafiantes y complejos a través de técnicas de preprocesamiento fue esencial para lograr resultados precisos. La limpieza y transformación efectiva de los datos directamente influyó en la calidad de los modelos finales.

La identificación cuidadosa de características relevantes para predecir el incumplimiento crediticio fue crucial. La selección precisa de características contribuyó a mejorar la precisión de los modelos y reducir la complejidad.

La presencia de un desequilibrio entre las clases de la variable objetivo presentó un desafío adicional. La implementación de técnicas específicas para manejar clases desequilibradas, como el muestreo estratificado o el ajuste de pesos, fue crucial para obtener predicciones más precisas y evitar sesgos.

La evaluación y prueba de varios modelos supervisados permitió determinar cuál era el más adecuado para el problema en cuestión. Identificar modelos con un mejor desempeño en términos de precisión y capacidad para abordar desafíos específicos de los datos fue un paso fundamental.

Además de la obtención de predicciones precisas, la comprensión y explicación de los factores clave que influyen en el incumplimiento crediticio resultaron fundamentales. Esto permitió proporcionar información valiosa a los interesados y tomar decisiones informadas en futuras estrategias de gestión de riesgos crediticios.

El proyecto destacó la importancia del aprendizaje continuo. La exploración de nuevas técnicas, la experimentación con diferentes enfoques y el mantenimiento de la actualización con los avances en el campo del aprendizaje automático fueron aspectos clave para el crecimiento profesional y la mejora continua.